Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko
Koroška cesta 46
2000 Maribor, Slovenija

Projektna naloga

Pri predmetu MOS

Avtor: Milan Djuric

Smer študija: ITK

Študijsko leto: 2. Letnik

1 Naive-Bayes klasifikator

Naivni Bayesovi klasifikatorji so vrsta algoritma za nadzorovano učenje, ki se uporablja za naloge klasifikacije. Imenujejo se "naivni", ker predpostavljajo, **da so vse funkcije v naboru podatkov neodvisne druga od druge**, kar ni vedno tako v podatkih iz resničnega sveta.

1.1 Opis

Kljub tej predpostavki lahko naivni **Bayesovi klasifikatorji** še vedno dobro delujejo v mnogih primerih, zlasti če so funkcije razmeroma neodvisne. Pogosto se uporabljajo pri nalogah klasifikacije besedila, kot je filtriranje neželene pošte, pa tudi na drugih področjih strojnega učenja in podatkovnega rudarjenja.

Osnovna ideja za naivnim **Bayesovim klasifikatorjem** je izračunati verjetnost, da dana podatkovna točka pripada vsakemu od možnih razredov, na podlagi značilnosti podatkovne točke in verjetnosti, da se te značilnosti pojavijo v vsakem razredu. Razred z največjo verjetnostjo je nato izbran kot napovedani razred za podatkovno točko.

Za usposabljanje naivnega **Bayesovega klasifikatorja** morate imeti označen nabor podatkov, kjer je znan razred vsake podatkovne točke. Klasifikator bo uporabil te podatke za usposabljanje za oceno verjetnosti vsake značilnosti, ki se pojavi v vsakem razredu, in bo uporabil te verjetnosti za napovedovanje novih, še nevidenih podatkov.

1.2 Osnovne značilnosti algoritma

Osnovne znacilnosti algoritma :

1. So enostavni za implementacijo in hitri za usposabljanje, zaradi česar so uporabni za velike nabore podatkov ali ko je treba usposabljanje izvesti hitro.

2. Obvladajo lahko veliko število funkcij in lahko delujejo z diskretnimi ali neprekinjenimi funkcijami.

3. Lahko obravnavajo manjkajoče podatke, če manjkajoči podatki niso preveč razširjeni.

4. Odporni so na prekomerno opremljanje, kar pomeni, da na splošno ne delujejo tako dobro kot bolj zapleteni modeli na zelo majhnih naborih podatkov, vendar se dobro obnesejo na večjih naborih podatkov in lahko posplošujejo nove podatke bolje kot bolj zapleteni modeli.

5. Lahko obdelujejo podatke s hrupom, saj predpostavka o neodvisnosti med funkcijami pomeni, da prisotnost šuma v nekaterih funkcijah ne bo bistveno vplivala na napovedi klasifikatorja.

6. So široko uporabljeni in dobro razumljeni, kar pomeni, da je na voljo ogromno virov in orodij za delo z njimi.

2 Podatkovna zbirka

Ime nabora podatkov, ki sem ga izbral, je **social network ads**.

2.1 Vsebina podatkovne zbirke

Vsebina podatkovne zbirke je res enostavna kjer imamo samo Age in Money.

Ciljna/glavna spremenljivka se imenuje **Class**. In določa, ali oseba ima raka ali ne. Uporablja se v formatu 0**/1** ali **true**/**false**.

3 Predstavitev rezultatov

Rezultati mojih meritev so bili narejeni v programskem jeziku Python, kjer sem v program implementiral naivno-bayas klasifikacijo brez uporabe že implementiranih algoritmov. Rezultati, ki sem jih dobil, so spodaj.

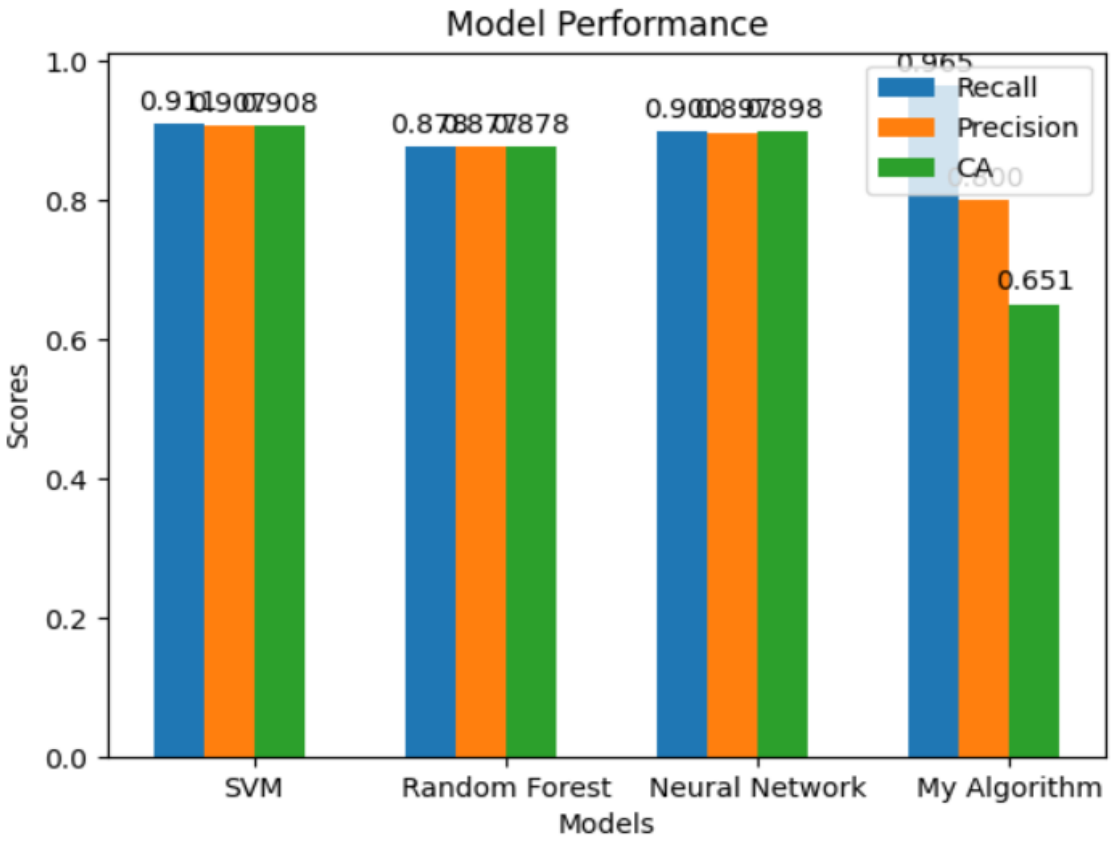
```
{'TP': 36, 'FP': 1, 'FN': 15, 'TN': 28}

accuracy: 0.8
sensitivity: 0.7058823529411765
specificity: 0.9655172413793104
precision: 0.972972972972973
{'accuracy': 0.8, 'sensitivity': 0.7058823529411765, 'specificity': 0.9655172413793104, 'precision': 0.972972972972973}
```

1. REZULAT MERITVE

3.1 Tabelarična predstavitev rezultatov

Model	Accuracy	CA	F1	Precision	Recall
SVM	0.9460851926977688	0.9075	0.9082745648810355	0.9110308348153573	0.9075
Random Forest	0.923556457065585	0.8775	0.8777647070702992	0.8781330437580438	0.8775
Neural Network	0.9543340094658553	0.8975	0.8981219834578336	0.899689182768451	0.8975
My Algorithm	0.8	0.8	N/A	0.972972972972973	0.7058823529411765

3.2 Grafična predstavitev rezultatov



3.3 Interpretacija rezultatov

V ovom primeru se najboljse pokazal SVM in Neural Network kot pricakovano. Moj algoritam se ni najboljse pokazal apmak mislim da je razlog to da je dataset zelo mali (400 uzoraka) in mislim da na test data 20% ni dovolj samo 50 da bi imeli dobar rezultat, zaradi tega mislim da je slab rezultat.

4 Zaključek

Vsaka klasifikacija je dosegla enak dobar rezultat za vse, kar smo iskali, občutljivost, specifičnost, priklic, natančnost in točnost.